Detecting False Identity through Behavioural Patterns

Shen, Qiang; Boongoen, Tossapon

Detecting False Identity through Behavioural Patterns

Tossapon Boongoen and Qiang Shen

Department of Computer Science, Aberystwyth University, UK
{tsb,qqs}@aber.ac.uk

Keywords: False identity, behavioural pattern, link analysis.

Abstract

Combating identity fraud is prominent and urgent since false identity has become the common denominator of all serious crime. Typical approaches to detecting false identity rely on the similarity measure of text-based identity attributes, which are usually not applicable to falsely-defined and unknown identities. This paper presents a novel link-based approach that can efficiently overcome such barrier. Its experimental evaluation against well-known link-oriented and text-based methods significantly indicates the great potential towards an effective verification system.

1 Introduction

False identity has become the common denominator of all serious crime such as mafia trafficking, fraud and money laundering. Holders of false identity all commonly determine to avoid accountability and traces for law enforcement authority. In essence, this offence is intentionally committed with a view to perpetrating another crime from the most trivial to the most dreadful imaginable. Organized criminals make use of counterfeit identity to cover up illicit activities and illicitly gained capital. Particularly in the case of terrorism, it is widely utilized to provide financial and logistical support to terrorist networks that have set up and encourage criminal activities to undermine civil society. Tracking and preventing terrorist activities undoubtedly require authentic identification of criminals and terrorists who typically possess multiple fraud and deceptive names, addresses, bank accounts, telephone numbers and email accounts.

With present high-quality off-the-shelf equipment, it is now effortless to obtain false identity documents. Conversely, it requires a great deal of time and experience to distinguish between genuine and forged copies. Usually, it is not feasible for a common person to recognize ten or fifteen security features presented in a document in a short period. However, a successful detection can prevent the revolting consequence like that of shocking September-11 terrorist attacks.

In particular to this tragedy, US authorities seriously failed to discover the use of false identities by nineteen terrorists, who were all able to enter the United States without any problem, in the very morning of the attacks. Most of them typically possess several dates of birth and multiple aliases. For instance, Mohamed Atta, alleged ringleader of the September 11 attacks, has exploited eight different aliases of Mehan Atta, Mohammad El Amir, Muhammad Atta, Mohamed El Sayed, Mohamed Elsayed, Muhammad Al Amir Awag Al Sayyid Atta, and Muhammad Al Amir Awad Al Sayad. To such extend, typical identity verification systems that rely solely on the inexact search of textual attributes would fail drastically to disclose unconventional truth.

The aforementioned dilemma may be overcome through link analysis, which seeks to discover knowledge based on the relationships in data about people, places, things, and events. Recently, this technique is employed by Argentine intelligence organizations to analyzing Iranian-Embassy telephone records in such a way to make a circumstantial case that the Iranian Embassy had been involved in the July 18, 1994, terror bombing of a Jewish community centre [6].

In light of such intelligent practice, this paper presents a novel link-based approach to detecting false identity based on the quantitative property of paths connecting any two particular identities. Its applicability and performance are experimentally evaluated against state-of-art link-based and text-based methods over a terrorism-related dataset.

2 Link Analysis

Intuitively, despite using several distinct false identities, each terrorist normally exhibits unique relations with other entities involving in legitimate activity found in any open or modern society, making extensive use of the Internet, mobile phones, public transportation, and financial systems. Inspired by this observation, false identity may be detected using link-based similarity measures, which, as indicated above, have proven effective for intelligence data analysis and also for identifying similar problems in the Internet and publication domains.

2.1 Network of Linked Entities

In order to disclose the possibility of false identity, a link network similar to that presented in Figure 1 is initially derived from collected intelligence data entailing activities of suspected identities (where each number represents the frequency that a pair of entities relate). Formally, the link network is modelled as an undirected graph $G = (V, E)$ in which entities are represented as vertices V and their relations are denoted as edges E.

Given $v_i, v_j \in V$, the similarity of two identities v_i and v_j can be estimated based on the cardinality measure of shared neighbours, which is the common intuition of several link-based similarity algorithms, for instance, *Co-citation* [7], *SimRank* [3] and *PageSim* [4]. The first two approaches were invented to find similar scientific papers using their citation relationship. In a different domain, the *PageSim* was developed to capture similar web pages based on associations implied by their hyperlinks. It is noteworthy that this algorithm explicitly uses the page ranking scheme, *PageRank* [5], of the Google[1] search engine.

With link-based similarity measures, for any two identities in a link network, *the higher their similarity is, the greater likelihood of forged identity becomes*. According to Figure 1, identity *B* is more similar to identity *A* than *C*: *B* and *A* are linked to two identical entities, while *B* and *C* share only one

[1] www.google.com

entity that is the house they live in. Hence, *A* and *B* are more likely to be false identities, comparing to *C*.

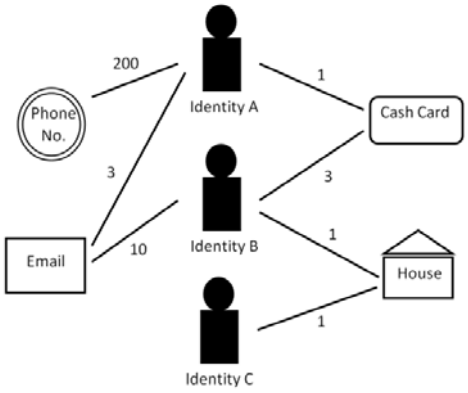

Figure 1: Example of a link network of intelligence data.

2.2 Uniqueness of Link Pattern

Despite their general applicability and decent performance, existing algorithms evaluate the similarity based entirely on the link structure, without exploiting link properties. In fact, the statistical information of a link (i.e. link frequency shown in Figure 1) can be employed to assess its uniqueness. This additional measure may help crystallizing the assessment of link-based similarity.

To illustrate the underlying concept of link uniqueness, the similarity of entities a and b within two different neighbouring contexts is elaborated. As presented in Figure 2, the link pattern $\{a, c, b\}$ is considered to be more unique, comparing to the pattern $\{a, d, b\}$. This conclusion is emphasized on the fact that links from entities $\{a, b\}$ to c equal to $3/3$ of the entire links between c and any entity, while the ratio for links from entities $\{a, b\}$ to d is $4/10$. Accordingly, the link pattern $\{a, c, b\}$ provides more confident context for justifying the similarity of a and b.

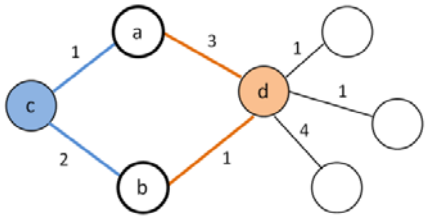

Figure 2: Uniqueness of link patterns $\{a, c, b\}$ and $\{a, d, b\}$.

2.3 Path-based Similarity Algorithm

Like *SimRank* and *PageSim* methods, the new path-based algorithm proposed here, called *Connected-Path*, takes into account the neighbouring context larger than the adjacent span previously discussed. A path between entities i and j is presented as $path(i,j)$, denoting a sequence of unique vertices $v_i, v_1, \dots, v_n, v_j$ such that edges $E_{v_i v_1}, E_{v_1 v_2}, \dots, E_{v_n v_j} \in E$. The length of such a path $L(path(i,j))$ is n, i.e. the number of vertices in the path excluding the two ends, v_i and v_j. In addition, a set of paths connecting i and j is denoted as $PATH(i,j)$.

Essentially, the similarity of entities i and j is determined by the uniqueness of paths, whose length ranges from *1* to r, between them. This measure can be formally defined as:

$$Connected-Path(i,j,r) = \sum_{p \in PATH(i,j), L(p) \le r} \frac{U(p)}{L(p)} \quad (1)$$

where $U(path(i,j))$ is the uniqueness of the path $path(i,j)$, which can be calculated using the following equation. Note that the uniqueness of a path is fractioned by its length, as longer paths are intuitively considered to be less informative comparing to shorter ones.

$$U(path(i,j)) = \prod_{v_x \in path(i,j), v_x \neq \{v_i, v_j\}} uScore(v_x) \quad (2)$$

where $uScore(v_x)$ is the uniqueness score measured at the vertex v_x that can be simply estimated as:

$$uScore(v_x) = \frac{|E_{v_x v_{x-1}}| + |E_{v_x v_{x+1}}|}{\sum |E_{v_x v_g}|} \quad (3)$$

An edge between the vertex v_x and any other vertex v_g in a network is denoted as $E_{v_x v_g}$, while $E_{v_x v_{x-1}}$ and $E_{v_x v_{x+1}}$ represent edges from v_x to v_{x-1} and to v_{x+1} respectively, such that $v_{x-1}, v_{x+1} \in path(i,j)$ and they are adjacent to v_x. In addition, $|E_{mn}|$ is the occurrence frequency of the edge between vertices m and n.

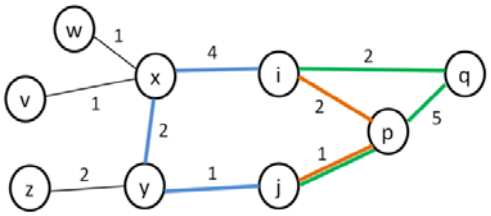

Figure 3: Paths between entities i and j.

According to the link network shown in Figure 3, there are two paths of length 2 ($p_1 = \{i, x, y, j\}, p_2 = \{i, q, p, j\}$) and one path of length 1 ($p_3 = \{i, p, j\}$) between entities i and j. The uniqueness of the path p_3 is estimated as:

$$U(p_3) = uScore(v_p) = \frac{2+1}{2+1+5} = 0.375$$

Similarly, the uniqueness of the longer paths p_1 and p_2 can be derived as follows:

$$U(p_1) = uScore(v_x) \times uScore(v_y)$$
$$= \frac{2+4}{2+4+1+1} \times \frac{2+1}{2+1+2} = 0.450$$
$$U(p_2) = uScore(v_q) \times uScore(v_p)$$
$$= \frac{2+5}{2+5} \times \frac{5+1}{5+1+2} = 0.750$$

According to Equation 1,

$$Connected-Path(i,j,2) = \frac{U(p_1)}{L(p_1)} + \frac{U(p_2)}{L(p_2)} + \frac{U(p_3)}{L(p_3)}$$
$$= \frac{0.45}{2} + \frac{0.75}{2} + \frac{0.375}{1} = 0.975$$

The similarity measure of entities i and j is obtained by the following normalization, where $Connected-Path(r)_{max}$ is the maximum estimate of $Connected-Path$ between any two entities within a link network, based on paths whose length range between 1 and r.

$$S_{Connected-Path}(i,j,r) = \frac{Connected-Path(i,j,r)}{Connected-Path(r)_{max}} \quad (4)$$

Given $Connected-Path(2)_{max}$ of the network presented in Figure 3 equals to 5, the path-based similarity of entities i and j is:

$$S_{Connected-Path}(i,j,2) = \frac{0.975}{5} = 0.195$$

By employing this terminology in false identity detection, the higher the measure of $S_{Connected-Path}(i,j,r)$, the greater the possibility that entities i and j involved in an identity fraud. It is noteworthy that longer link paths (i.e. higher r) make the overall measure more refined and robust, but at the cost of greater computational requirement. Figure 4 shows the intuition of path spectral, with various values of r, exploited in this estimation of similarity.

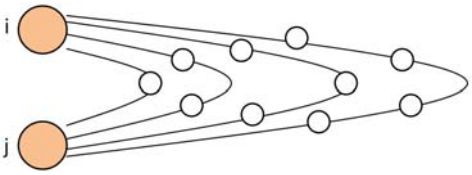

Figure 4: Spectral of link paths ($r = 1 \ldots 4$) between entities i and j.

3 Performance Evaluation

In order to evaluate the performance of the Connected-Path approach, its similarity estimates are compared with those derived by other link-based algorithms and the traditional textual similarity measure, for the task of finding aliases in the *Terrorist* dataset [1]. This is a link dataset manually extracted from web pages and news stories related to terrorism. Each entity presented in this link network is a name of person, place and organization, while a link denotes an association between objects through reported events. Statistically, this network contains 4088 entities, 5581 links and 919 alias pairs (i.e. false identities). Figure 5 shows an example of this link network in which names *Bin laden* and *Abu abdallah* truly refer to the same real-world person.

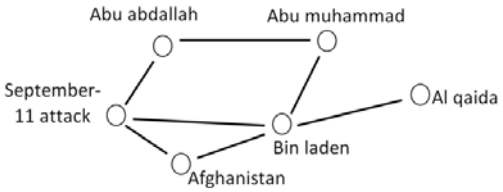

Figure 5: An example of *Terrorist* link dataset.

3.1 Performance of Connected-Path Algorithm

The Jaro [2] string similarity measure, extensively utilized in database and record linkage domains, is employed here to illustrate the effectiveness of text-based methods and also to set the base-line performance for link-based approaches to overcome. Table 1 depicts the amount of alias pairs disclosed by each method over several sets of k entity pairs with the k highest similarity values (where *k* = 200, 400, 600, 800 and 1000, respectively) in the studied network. Accordingly, the new Connected-Path algorithm consistently outperforms the text-based Jaro method over these five sets of top-k similar entity pairs. Similarly, other experimented link-based measures, PageSim and SimRank, disclose significantly lower numbers of alias pairs, especially the SimRank method that fails to discover any alias in the top-*600* entity pairs.

Method	*k* pairs with highest similarity values				
	200	**400**	**600**	**800**	**1000**
Connected-Path	52	81	136	170	193
PageSim	7	36	63	79	92
SimRank	0	0	0	1	2
Jaro	22	33	40	43	47

Table 1: Numbers of alias pairs discovered by each method.

It is noteworthy that the results of the Connected-Path method shown in Table 1 is achieved by considering paths with length up to 3 (i.e. $r = 3$). Essentially, its time complexity can be reduced by setting $r = 1$, in which case the number of disclosed aliases drops slightly, but it is still much larger than those of its counterparts (see statistical details Figure 6). This implies the efficient exploitation and flexibility of the Connected-Path algorithm in real-time applications.

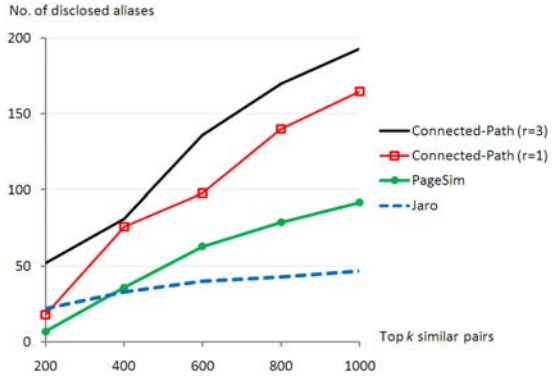

Figure 6: Performance of Connected-Path with $r = 3$ and $r = 1$.

3.2 Deficiency of Text-based Algorithm

The text-based similarity measures like Jaro perform well when aliased entities possess similar string content, caused likely by misspelling, data-entry or translation errors. For instance, the following two alias pairs are efficiently recovered with high Jaro scores of 0.9881 and 0.9556, respectively.

- *fahd bin abdallah bin khalid, fahd bin adballah bin khalid*
- *osama bin laden, usama bin laden*

In contrary, the Jaro measure fails to evaluate the similarity of several alias pairs each without overlapping textual content. For instance, the following alias pairs are not recovered by the Jaro method, each with Jaro score of *zero*.

- *ashraf refaat nabith henin, salem ali*
- *fahid mohammed ali msalam, usama al-kini*
- *fadil abdallah muhamad, harun fazul*
- *bin laden, the prince*
- *bin laden, the emir*
- *abu mohammed nur al-deen, the doctor*
- *abu anis, moustafa ali elbishy*

This crucial limitation greatly constrains the application of text-based methods within identity verification systems.

Specifically to the Terrorist dataset, there are 183 alias pairs (out of 919 pairs) that will never be discovered using the Jaro measure (see Table 2 for details).

Jaro similarity score (J)	Number of alias pairs
$J \geq 0.8$	104
$0.8 > J \geq 0.6$	224
$0.6 > J \geq 0.4$	381
$0.4 > J \geq 0.2$	27
$0.2 > J > 0.0$	0
J = 0.0	**183**

Table 2: Number of alias pairs in Terrorist dataset categorized into sets of different Jaro score intervals.

This barrier is partially resolved as 92 out of these 183 problematic cases can be recognized using the Connected-Path method. Examples (each marked with Connected-Path score, which is, however, very small due to the fact that most paths connecting these names are either long or with extremely low uniqueness scores) of such case are:

- *ashraf refaat nabith henin, salem ali* (0.077)
- *fahid mohammed ali msalam, usama al-kini* (0.072)
- *fadil abdallah muhamad, harun fazul* (0.046)
- *bin laden, the prince* (0.003)

Henceforth, identity verification systems may gain essential benefit by including the Connected-Path similarity measure, in addition to traditional content comparisons of text-based identity attributes.

3.3 Performance of Combined Method

Despite its prescribed deficiency, the Jaro algorithm can be used to purify the similarity measured generated initially by the Connected-Path method. Figure 7 shows the process model of the combined approach.

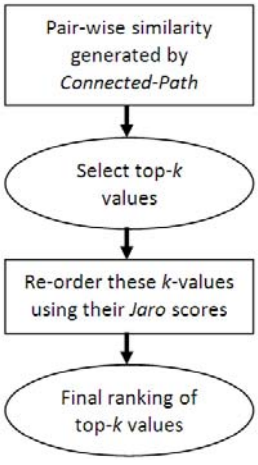

Figure 7: Sequential process model of the combined method.

With the Terrorist dataset and $k = 1000$, Figure 8 presents the improvement made by this simple combined method over the Connected-Path algorithm.

4 Conclusion

This paper presents a novel link-based method to detecting false identity which is a crucial problem in intelligence data analysis. Its performance over terrorism-related dataset is far more superior to text-based and other well-known link-based approaches.

Despite this achievement, the application of qualitative link properties such as meanings or link labels is to be further examined in order to refine the similarity estimation. Moreover, the proposed link-based similarity measure is to be exploited for resolving identities and aggregating relevant scenarios in the environment of intelligence data analysis.

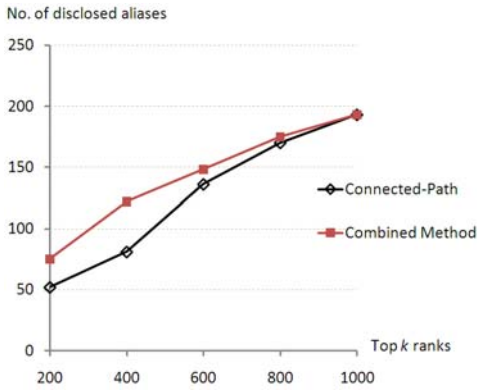

Figure 8: Improved performance of the combined method.

Acknowledgements

This work is sponsored by the UK EPSRC grant EP/D057086. The authors are grateful to the members of the project team for their contribution, but will take full responsibility for the views expressed in this paper.

References

[1] P. Hsiung, A. Moore, D. Neill and J. Schneider. "Alias Detection in Link Data Sets", *Proceedings of International Conference on Intelligence Analysis*. 2005.

[2] M.A. Jaro. "Probabilistic linkage of large public health data files", *Statistics in Medicine*, Vol. 14, pp. 491-498. 1995.

[3] G. Jeh and J. Widom. "SimRank: A Measure of Structural-Context Similarity", *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.538-543. 2002.

[4] Z. Lin, I. King and M.R. Lyu. "PageSim: A Novel Link-Based Similarity Measure for the World Wide Web", *Proceedings of International Conference on Web Intelligence*, pp. 687-693. 2006.

[5] L. Page and S. Brin. "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, Vol. 33, pp. 107-117. 1998.

[6] G. Porter. "Crying (Iranian) wolf in Argentina", *Asia Times Online* (www.atimes.com), Jan 25, 2008.

[7] H.G. Small. "Co-citation in the scientific literature: A new measure of relationship between two documents", *Journal of the American Society for Information Science*, Vol. 24, No. 4, pp. 265-269. 1973.